Econometric Analysis of Startups & Their Success

By: Dillion Cottrill

West Virginia University

December 9th, 2020

Introduction

In this study, I am doing an analysis on various factors which contribute to the success of startups. In particular, we are interested in their status as of the time of the data acquisition. We will determine the success of a company as being “Acquired”. Whether or not a company was actually acquired by another is irrelevant in this study, as if a company is still operating on its own it is still accounted for as “Acquired”. If the company is “closed” this means that the company is shut down. We will take into account several data types in order to predict the outcome of a company, which include: type of company (ecommerce, web based, software development), funding(total), number of relationships, location(NY, TX, MA, CA), and finally number of milestones reached.

This analysis will be useful for anyone curious what factors will contribute to their success the most. There are a lot of start ups founded every year, and quite a bit of them fail. Using my analysis, I would hope that people could gain some insight into what is key for success, and what is not actually significant. There isn’t a clear consensus on how to become an entrepreneur, and so analyzing the factors is a great way to assist those who are unsure or wary of entering this field. A lot of people have great ideas, but because of poor execution, they fail to acquire funding, proper location, or partners to help create their vision. With this analysis, I hope to make some of the most important factors clearer, to help people, and myself, form clear goals when it comes to creating a business.

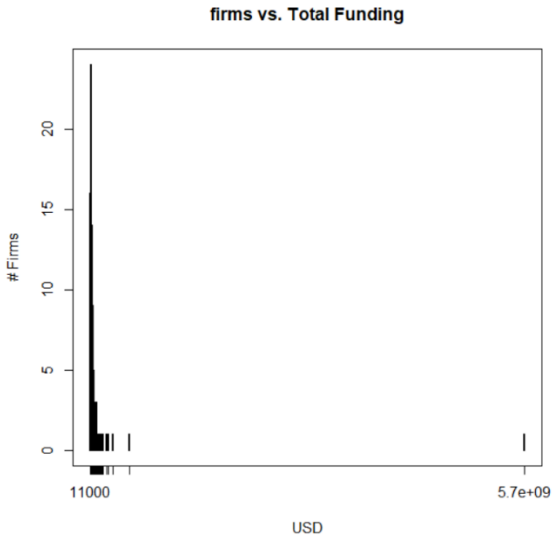
I expect the most significant and telling factors will be funding and relationships. These are key factors which allow for people to fall back in the case of a risk which wasn’t properly accounted for. After all, money trees are the perfect places for shade. Unfortunately, even with a

lot of funding, we can find firms making the same mistakes over and over again, and so it will be interesting to see just how significant this variable will be. Relationships are key, as they allow firms to adequately leverage their goods and services to other businesses, partners, and third parties which are interested in either advertising or consuming their products. Relationships are truly how the business world works, as without them, you have no one to work with and to make money with or from.

Data

My data comes from a Kaggle data set submission which was used in DPhi's data sprint #5. This was most likely a data set used for machine learning, and has quite a few data entries, just over 900. The data is from several states all over the United States. All in all, there are 48 columns of data to work with, though I will only be focusing on 11 of them, with one being an indication of success or failure, and the rest being used to indicate the probability of success or failure. Although it is quite a large data set, the columns I selected are mostly categorical, and none of them are missing values, which I find incredible.

I wanted to look at the distribution of some of the data, so I used Stargazer to output some Latex code to create a table on the following page (Table 1). After removing one extremely large company with massive start up funding, we had some reasonable estimates.



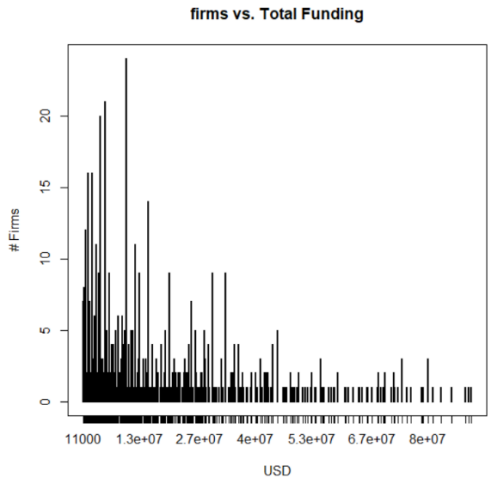


Figure 1: Before removing Outliers

Figure 2: illustration of firms with $11,000-$90,000,000 in funds

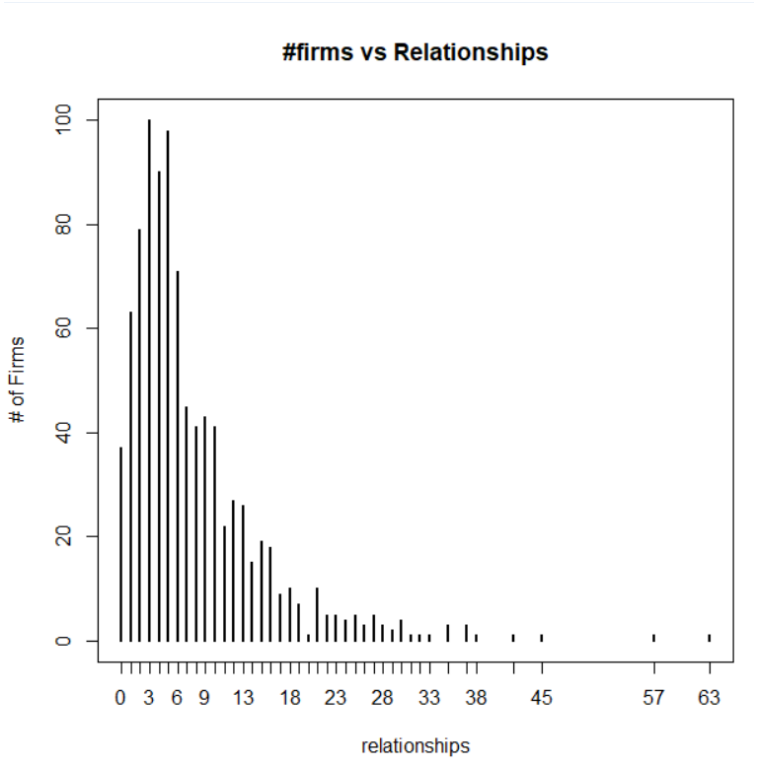
Table 1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
relationships	922	7.698	7.260	0	3	10	63
funding_total_usd	922	19,265,107.000	31,613,044.000	11,000	2,712,500	24,652,635	510,000,000
milestones	922	1.842	1.323	0	1	3	8
is_CA	922	0.528	0.499	0	0	1	1
is_NY	922	0.115	0.319	0	0	0	1
is_MA	922	0.090	0.286	0	0	0	1
is_TX	922	0.046	0.209	0	0	0	1
is_software	922	0.166	0.372	0	0	0	1
is_web	922	0.156	0.363	0	0	0	1
is_ecommerce	922	0.027	0.163	0	0	0	1
status	922	0.646	0.478	0	0	1	1

I removed the data point at $5.7e+9 because it was simply so much larger than any other data point, and I don't feel it is reflective of the average entrepreneur. There are many data sets

which are large and considered outliers if we use the quartile range rule, but I will leave them in the chance they reflect well in the model.

The same can be said for relationships, as there are many values above the standard deviation. I have plotted some information on that relationship below.



Empirical Strategy

In this section, we will discuss the reasoning behind the methodology I have chosen for the analysis of this project. I would like to investigate several linear models which attempt to predict the probability of the binary variable which represents current status.

The first will be to create a linear probability model using the funding variable alone.

$$status = fund * X_1 + \epsilon$$

We use the normal method of calling the linear model function with status and funding_total_usd. I used this very basic model because I was curious of the significance of funding on success. This model isn't quite good enough to develop any solid ideas about whether you will be successful, because there is no doubt many terms nested into the error term for the regression when we run this simple case. We know that funding is not the only determining factor of success, as many other factors will contribute to the result of a company. It will be useful however, to examine this simple relationship as a preliminary step:

My next topic of interest will be how funding and relationships interact in different areas, like Texas, California, New York, and Boston.

$$status = \beta_1 fund + \beta_2 rel + \beta_{fr} rel * fund + \beta_3 MA + \beta_4 NY + \beta_5 CA + \epsilon$$

In changing the amount of funding, while holding the number of relationships constant, how will our prediction of the interaction coefficient change? I predict that there will be some positive value for the coefficient. It would make sense that the more money you have, the more people you are working with and selling to.

I would like to investigate the significance of milestones in predicting success, especially within the fields of ecommerce, web, and software.

$$status = \beta_1 mile + \beta_2 ecom + \beta_3 web + \beta_4 soft + \epsilon$$

These are big fields, with extremely competitive markets and high profits. It seems logical to assume that this may be the single most important factor when trying to predict success. A technical company which consistently is accomplishing its goals will no doubt find the funding it needs to continue, whether it's based in Tennessee or Silicon Valley. As people

hear about a high preforming company, they will no doubt be interested in investing into the company.

After looking over the results of the previous analyses, I wanted to conduct one more model:

$$status = \beta_{mr} mil * rel + \beta_1 mil + \beta_2 rel + \beta_3 MA + \beta_4 ecom + \epsilon$$

I mostly composed this model because it seems that the terms described here consistently had extremely small p values, indicating it to be improbable of observations far from expected values, and large contributions to the probability of success. This model will be the most exciting to pick apart, as it has some values which seem to be the most useful when attempting to decide what kind of business to create.

Analysis

For the first model, we get the following statistics:

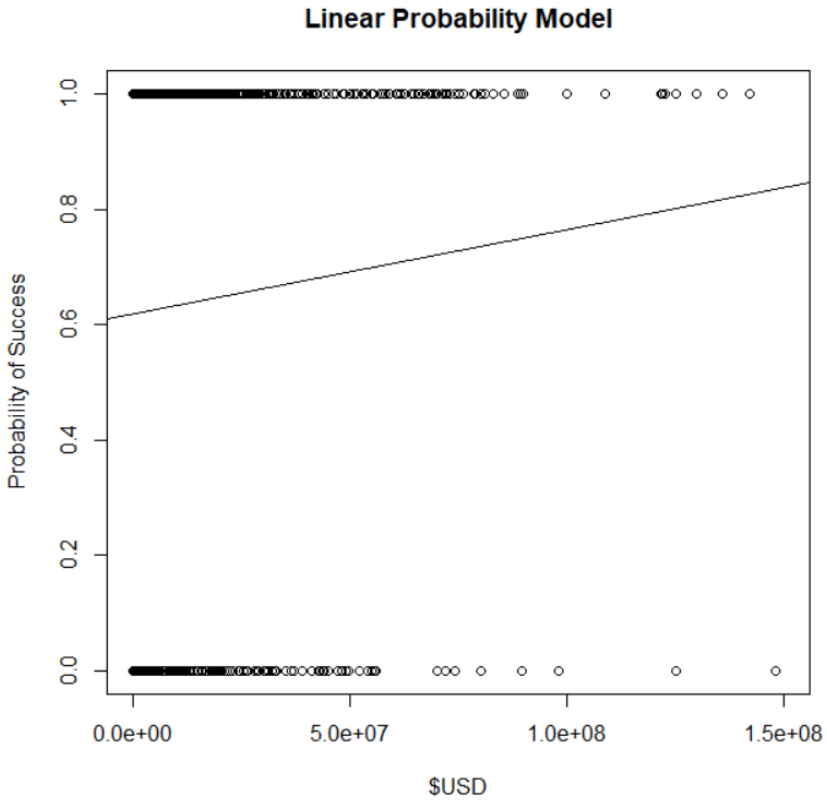
```
Call:
lm(formula = status ~ funding_total_usd, data = entre)

Residuals:
    Min      1Q  Median      3Q     Max
-1.3677 -0.6200  0.3398  0.3681  0.3818

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.181e-01  1.837e-02   33.64  < 2e-16 ***
funding_total_usd 1.470e-09  4.965e-10    2.96  0.00315 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4763 on 920 degrees of freedom
Multiple R-squared:  0.009435,  Adjusted R-squared:  0.008359
F-statistic: 8.763 on 1 and 920 DF,  p-value: 0.003152
```

Here we have ran a simple linear regression on the status vs. the funding. Here is a plot of the linear probability model, with the line of best fit drawn:



At the intercept, or at zero funding, we have a probability of success of around 60%. This means that even around low levels of funding, we still have above a 60% chance of success to succeed. Of course, it's unlikely a business would maintain itself for long with zero funding. Regardless, status seems to be correlated with funding weakly. If we run a correlation test, we find that we have a value of .097, a weak but positive correlation.

There is also the problem of the obvious heteroskedasticity, which we will confirm using a Breusch-Pagan test:

```
data:  L0
BP = 5.2375, df = 1, p-value = 0.02211
```

The p-value is below the threshold of .05 and we reject the hypothesis of homoskedasticity and assume heteroskedasticity. Thus, we run coefficient test with some resources I found on how to correct for heteroskedasticity[1]:

```
> coeftest(L0, vcov = vcovHC(L0, "HC1"))

t test of coefficients:

                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       6.1811e-01 2.2123e-02 27.9393  < 2e-16 ***
funding_total_usd 1.4698e-09 8.3430e-10  1.7617  0.07845 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These values are corrected for heteroskedasticity, and better reflect the standard errors we would see with it present. Our p-value is quite large, and funding doesn't seem to have as large of an impact as I expected. Let's analyze our next model.

The second model's summary, after corrected for heteroskedasticity using the same process of confirming via the bptest and using the robust standard errors, we get:

```
t test of coefficients:

                                  Estimate  Std. Error t value  Pr(>|t|)
(Intercept)                     3.8917e-01  3.4662e-02 11.2275 < 2.2e-16 ***
funding_total_usd               1.2042e-09  9.4125e-10  1.2793 0.2011071
relationships                   2.7255e-02  3.1224e-03  8.7291 < 2.2e-16 ***
is_CA                           6.1775e-02  3.2803e-02  1.8832 0.0599904 .
is_MA                           1.7152e-01  5.1911e-02  3.3041 0.0009898 ***
is_TX                          -2.7583e-02  8.2490e-02 -0.3344 0.7381676
funding_total_usd:relationships -1.0860e-10  4.4243e-11 -2.4547 0.0142868 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It seems that the most significant factors here are the relationships, is_MA variable, and the interaction variable between funding and relationships. In this model it seems that funding has little significance and doesn't seem to represent much in the outcome.

The third model seems to be a failure. With corrected homoscedasticity:

```
t test of coefficients:

              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.422664   0.029190 14.4797  < 2e-16 ***
milestones    0.125252   0.011029 11.3571  < 2e-16 ***
is_ecommerce -0.218138   0.091525 -2.3834  0.01736 *
is_web       -0.075174   0.040706 -1.8468  0.06510 .
is_software   0.064733   0.041584  1.5567  0.11989
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This highlights one important thing: milestones effect on status. That is the largest value we have thus far seen, with a change in ~.13, or 13% probability of success for each milestone. The p value is also exceedingly small, and this is an exciting result. Thus let us run another model, gathering up all of the values which seem to be the most interesting.

Model 4 is quite exciting:

```
t test of coefficients:

                           Estimate Std. Error t value  Pr(>|t|)
(Intercept)                0.2549222  0.0333887  7.6350 5.670e-14 ***
milestones                 0.1223238  0.0153777  7.9546 5.280e-15 ***
relationships              0.0375072  0.0045690  8.2090 7.533e-16 ***
is_MA                      0.1327393  0.0474798  2.7957  0.005287 **
is_ecommerce              -0.1940028  0.0817263 -2.3738  0.017811 *
milestones:relationships -0.0067603  0.0012616 -5.3587 1.061e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 4 states the same which model 3 stated; that with more milestones, you significantly increase your chances of success. Not only that, but the p value is small, we see a reasonable standard error, and the results are similar to the last regression. Relationships is also a factor with a small p value, which tells me that someone may be able to expect values like this.

It also has a fair standard error with respect to the estimate. Is_MA gives us values which seem to indicate that if you're in MA, you can automatically assume your chances for success are a fair deal higher.

Relationships has quite a bit of practical significance. The average value as stated above was around 8, and thus this has quite a bit of sway in the outcome.

Ecommerce seems practically insignificant, as well as statistically. I think this model could've done without.

The interaction value is small, though meaningful. We can obviously see some extremely small sway. I could imagine that one firm with too many relationships may be depending on others too much for certain things, but its honestly really difficult to say what exactly is happening to cause the interaction term to have such a small, but negative value. I would say this has low practical significance.

Discussion

I had no idea that there would be so much dependence upon relationships, milestones, and location. I originally thought that much of the importance would be on money, and that's just how I felt. In general, 2 of my 3 models predicted milestones as having a significant impact on the outcome of the regression, and I would like to infer that this seems to be the strongest indicator of success. It makes sense, because someone who can get stuff done will obviously make the money from doing so, the reputation, as well as the friends to go with it all.

As for the Massachusetts binary variable in the 3rd and 4th model, I think it is quite a good statistic in both. It has a very small p value, and has the strongest sway over the predictor.

We may be able to make policy from this, but I would say that it requires further research. I'm not entirely sure why a particular location would have so much sway over the success of a firm, but if we could figure out why then we could potentially create policy which reflects these conditions.

Conclusions

This was quite an exciting project, and I want to look at present day firms and see how they are doing in Massachusetts. I would also like to gauge what realistic milestones are, and how the information ripples through a business community. It was incredible to see the indication of these variables on the predictor. Another step in the future would be to figure out a good balance in terms of the number of relationships a firm has, and how two firms may cooperate to achieve goals.

To improve upon this study, I would consider including more data, sorting and presenting the data a bit better, and discussing what kind of milestones and relationships firms have with each other. A map of firm's connections as well as firms with mutual or rival interests would be fantastic to see as well. Entrepreneurship is an exciting field, and I look forward to learning more about it in the future.

References

[1] https://rstudio-pubs-

static.s3.amazonaws.com/187387_3ca34c107405427db0e0f01252b3fbdb.html